User-based topic, word, and sentiment analysis of Turkish tweets on platform X

Mehmet Yusuf BİRCAN[1], Ayşe ELDEM[1,*]

[1] Karamanoğlu Mehmetbey University Faculty of Engineering, Department of Computer Engineering, Karaman.

Geliş Tarihi (Received Date): 26.07.2025
Kabul Tarihi (Accepted Date):06.01.2026

Abstract

Social media platforms, especially X (Twitter), provide rich data sources for understanding social and individual trends. While existing studies generally focus on general data sets, analyses at the individual user level remain limited. This study aims to fill this gap by presenting a web-based system that extracts and analyzes data from specific users from X. The developed system collects tweets from the desired user using the web scraping technique and preprocesses this data with steps specific to the Turkish language. Then, it applies three basic analyses with Latent Dirichlet Allocation (LDA): topic modeling, sentiment analysis, and word cloud generation. The system visualizes the results of topic distributions, sentiment graphs, and word clouds through a user-friendly interface. This study presents an original tool for understanding individuals' interests, emotional states, and mindsets in more detail by providing an in-depth user-based perspective.

Keywords: *X, latent dirichlet allocation, text analysis, sentiment analysis, topic modeling, word cloud*

X platformunda üzerinde Türkçe tweetlerin kullanıcı bazlı konu, kelime ve duygu analizi

Öz

Sosyal medya platformları, özellikle X (Twitter), toplumsal ve bireysel eğilimleri anlamak için zengin veri kaynakları sunmaktadır. Mevcut çalışmalar genellikle genel veri kümelerine odaklanırken, bireysel kullanıcı düzeyindeki analizler sınırlıdır. Bu çalışma, X'teki belirli kullanıcılardan veri toplayan ve analiz eden web tabanlı bir sistem sunarak bu boşluğu doldurmayı amaçlamaktadır. Geliştirilen sistem, web kazıma tekniğini

Mehmet Yusuf BİRCAN, mehmet.bircan.2002@gmail.com, http://orcid.org/0009-0007-5728-3020
[*]Ayşe ELDEM, ayseeldem@kmu.edu.tr, http://orcid.org/0000-0002-5561-1568

kullanarak istenen kullanıcıdan tweet toplar ve bu verileri Türkçe diline özgü adımlarla ön işleme tabi tutar. Ardından, Latent Dirichlet Allocation (LDA) ile üç temel analiz uygular: konu modelleme, duygu analizi ve kelime bulutu oluşturma. Sistem, konu dağılımlarının, duygu grafiklerinin ve kelime bulutlarının sonuçlarını kullanıcı dostu bir arayüz aracılığıyla görselleştirir. Bu çalışma, derinlemesine kullanıcı tabanlı bir bakış açısı sunarak bireylerin ilgi alanlarını, duygusal durumlarını ve zihniyetlerini daha ayrıntılı olarak anlamak için özgün bir araç sunmaktadır.

Anahtar kelimeler: *X, latent dirichlet allocation, metin analizi, duygu analizi, konu modelleme, kelime bulutu*

1. Introduction

Data analysis has gained great importance today, especially with the spread of platforms such as social media [1]. These platforms, where people share their thoughts, feelings and experiences, have become huge data sources containing valuable information. However, this data is complex and scattered in its raw form. Thanks to data analysis, this raw data can be processed, made meaningful and used to make strategic decisions, identify trends and make predictions in different sectors [2]. For example, in the tourism sector [3], companies use data analysis to understand consumer behavior, develop their products and services and create more effective marketing campaigns for their target audiences. In the health sector [4], social media data is analyzed to track the spread of diseases, identify risk groups and shape public health policies. In the finance sector [5], data analysis tools are also used to measure investor sentiment, predict market trends and manage risks.

X (Twitter) is a general social media platform that provides data on all kinds of personal and social issues thanks to its large user base and the ability of users to share content freely [6]. Users share a wide range of content from their daily lives to their interests, from social events to their personal thoughts [7], making the platform a rich data source for individual research. For social media researchers, the big data provided by X plays an important role in examining users' behaviors, mindsets, and emotional states. In particular, it is a suitable platform for collecting and analyzing individuals' real-time reactions to current events, political views [8], or personal interests [9]. However, conducting sentiment analysis using large-scale data obtained from X is of great importance, especially in terms of determining users' tendencies.

In the literature study, it is noted that the articles investigated generally focus on general tendencies and that a user-based analysis approach is lacking. This study aims to provide an original perspective by addressing the aforementioned deficiency. The content of the application developed within the scope of this study is as follows.

- By analyzing the posts of the desired users on X, a user-based thematic and emotional content review was conducted.
- Thanks to this user-focused analysis, it is possible to reach a more detailed and personalized understanding by taking into account the individual characteristics, preferences and experiences of each user. This made the results of the study more effective and applicable.
- Not only the sentiment analysis of the users, but also the topics they talked about were examined.

- The data was pulled from X in real-time and analyzed directly with the relevant user data.
- Data collection, pre-processing, analysis and visualization steps were presented to the users through a single system.
- Thanks to the developed flask-based web interface, the analysis results obtained were conveyed to the users in an interactive manner.

The article is structured as follows. In the second section, literature analysis is provided, in the third section, the methodology of the developed system, data collection, data storage, analysis module, web interface and the technologies and libraries used are discussed. In the analysis module, data pre-processing steps and analysis techniques are presented in detail. In the fourth section, the findings obtained from the application interfaces and a sample user are presented and discussed. Finally, in the fifth section, the results are evaluated and suggestions are made for future studies.

2. Literature review

Within the scope of this study, studies conducted with the mentioned methods in the literature were examined and some examples of relevant studies are presented below.

Uzun's study aims to reveal how people felt about this event and how these feelings changed over time through the sentiment analysis of messages sent on X after the earthquake that occurred on February 6, 2023. The research emphasizes that sentiment analysis conducted on social media is an important tool in understanding human psychology in times of crisis [10]. Kartal examined the topic modeling of articles in TOJDE (The Turkish Online Journal of Distance Education) using the LDA algorithm. With this method, it was tried to determine which topics were covered in the journal, how these topics changed over the years and which topics were more popular, in his thesis [11]. In Taşbaş's study, word cloud analysis was used as a visual tool to reveal research trends in human-robot interaction. Word cloud analysis helped understand certain themes and trends by visualizing important words and expressions obtained from text data [12]. Günyaktı and Bursa found that sentiment analysis conducted on X data during the COVID-19 pandemic showed that positive perceptions towards teachers and healthcare workers were dominant, revealing that the general stance of society was based on appreciation and gratitude [13]. Seren and Altıntaş have addressed how tourism businesses can improve their brand perceptions by using customer comments on the internet using the sentiment analysis method. As a result of the research, it has been revealed how brand personality dimensions are shaped for tourism businesses and the emotional values that customers attribute to these businesses [14]. İlhan and Sağaltıcı performed sentiment analysis using X data. They classified tweets as positive and negative using Naive Bayes and Support Vector Machine methods [15]. Akgül et al. conducted sentiment analysis on tweets taken from a specific keyword. In this study, tweets were labeled as positive, negative and neutral using the n-gram and dictionary model [16]. Aydin and Hallac used the LDA algorithm to group news documents into 4 different classes [17].

These studies show how powerful text analysis techniques can produce results in different areas. The most important differences that distinguish this study from others are;

- While most of these studies focus on general events, large audiences or specific keywords, this study aims to make an original contribution to the literature by providing an in-depth analysis at the individual user level.
- The web interface, developed based on flask, provides a user-friendly experience.
- A data collection system integrated with open-source APIs for the X platform has been developed.

3. Material and method

In this study, a multi-stage methodology designed to analyze the shares of a specific user X was followed. In this section, the architecture of the developed system, the technologies used, the modular design and the data flow processes are presented in detail. The system is a web-based application that provides an end-to-end analysis infrastructure to produce meaningful insights from X data. The workflow of the system consists of four modules, namely Data Collection, Data Preprocessing, Analysis Engine and Web Interface, as shown in Figure 1. These components are designed with a modular structure and integrated with each other.

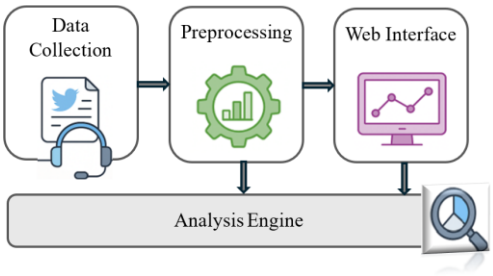


Figure 1. X Data Extraction and Analysis

3.1. Data collection module

The first step of the analysis process is to collect tweets of the targeted user X. In this study, web scraping technique is used to collect tweets of a user in X. This technique is used to access information quickly and effectively rather than manually collecting large data sets. Web scraping is one of the frequently preferred methods, especially for large-scale data analysis, data mining and various research purposes [18]. The system in Figure 2 retrieves tweet texts, sharing dates, and other metadata that are publicly available from the user's profile page. At this stage, it is essential to act in accordance with X's API usage policies and legal regulations. This module performs data retrieval from X using the twikit library. Tweets from the X account specified by the user are retrieved asynchronously and saved in JSON format. Thanks to cookie-based session management, API limitations are overcome and more secure access is provided.

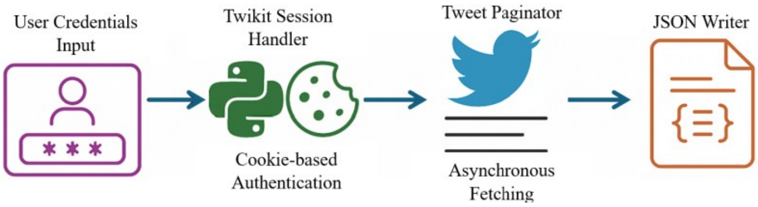


Figure 2. X Data Collection Module

The data captured and analyzed in the system is managed on a file basis, as shown in Figure 3. The captured tweets are stored in json format under the tweet archives/ folder. The analysis outputs are stored in the analysis_sonucları/ folder under the lda/, sentiment/ and wordcloud/ folders in CSV, PNG and HTML formats suitable for the file.

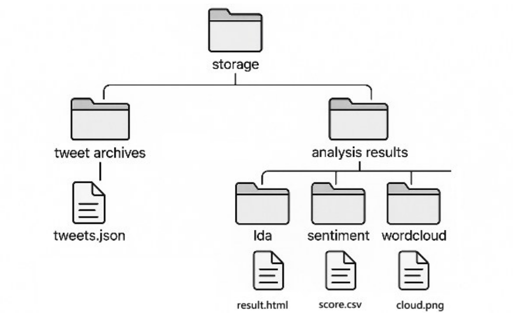
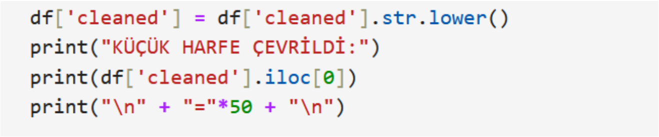


Figure 3. Data Storage Diagram

3.2. Data preprocessing module

It is a fundamental step in making texts suitable for modeling. Raw texts are often noisy and chaotic, which can make it difficult for algorithms to work correctly. The goal of data preprocessing is to make raw data suitable for modeling, resulting in more accurate and reliable results. Raw data often contains errors, omissions, inconsistencies, noise, and redundancies [19]. The raw text data collected within the scope of this study is subjected to a series of pre-processing steps in order for the analysis algorithms to work correctly. These steps were carefully selected considering the structure of Turkish texts:

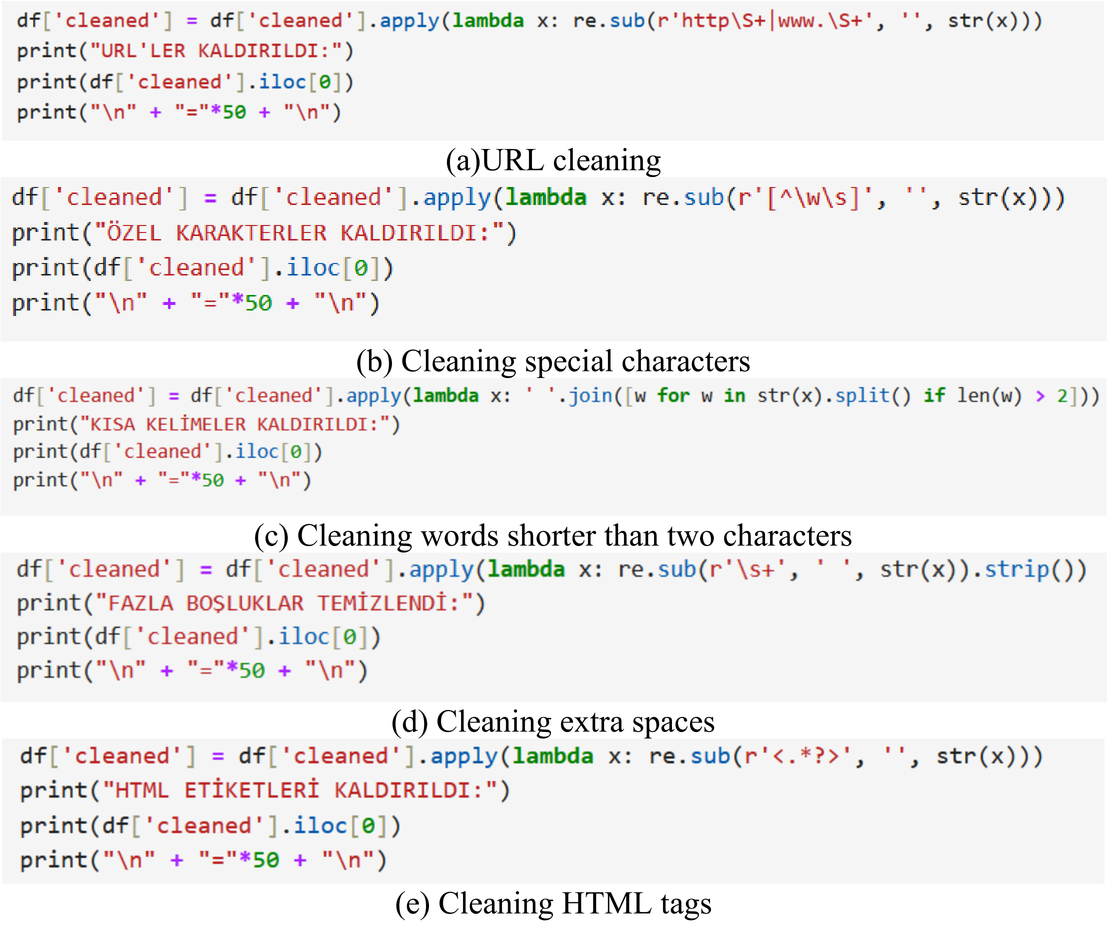
Convert to Lowercase: All letters in the text are converted to lowercase. The relevant code block is given in Figure 4.



```
df['cleaned'] = df['cleaned'].str.lower()
print("KÜÇÜK HARFE ÇEVRİLDİ:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

Figure 4. Data Preprocessing - Converting to lower case

Noise Cleaning: Components with no semantic value, such as URL addresses, hashtag symbols (#), and retweet notifications ("RT"), which are frequently found in tweet texts, are removed. The relevant code block is detailed in Figure 5.



```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'http\S+|www.\S+', '', str(x)))
print("URL'LER KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

(a)URL cleaning

```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'[^\w\s]', '', str(x)))
print("ÖZEL KARAKTERLER KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

(b) Cleaning special characters

```
df['cleaned'] = df['cleaned'].apply(lambda x: ' '.join([w for w in str(x).split() if len(w) > 2]))
print("KISA KELİMELER KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

(c) Cleaning words shorter than two characters

```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'\s+', ' ', str(x)).strip())
print("FAZLA BOŞLUKLAR TEMİZLENDİ:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

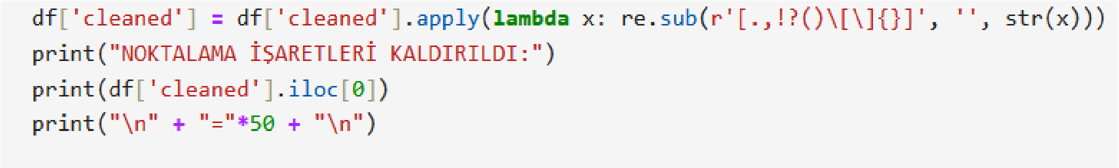
(d) Cleaning extra spaces

```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'<.*?>', '', str(x)))
print("HTML ETİKETLERİ KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

(e) Cleaning HTML tags

Figure 5. Data Preprocessing – Noise Cleaning

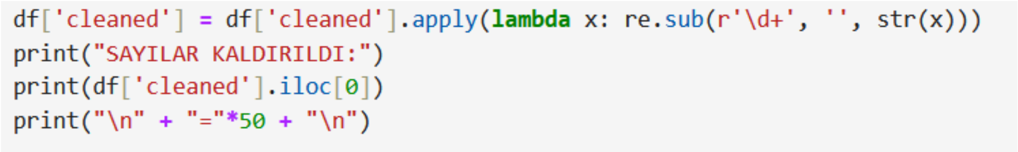
Punctuation Marks and Removal: Punctuation marks such as end-of-sentence marks, commas, and parentheses are cleaned from the text. The relevant code block is detailed in Figure 6.



```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'[.,!?()\[\]{}]', '', str(x)))
print("NOKTALAMA İŞARETLERİ KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

Figure 6. Data Preprocessing – Punctuation Marks Removal

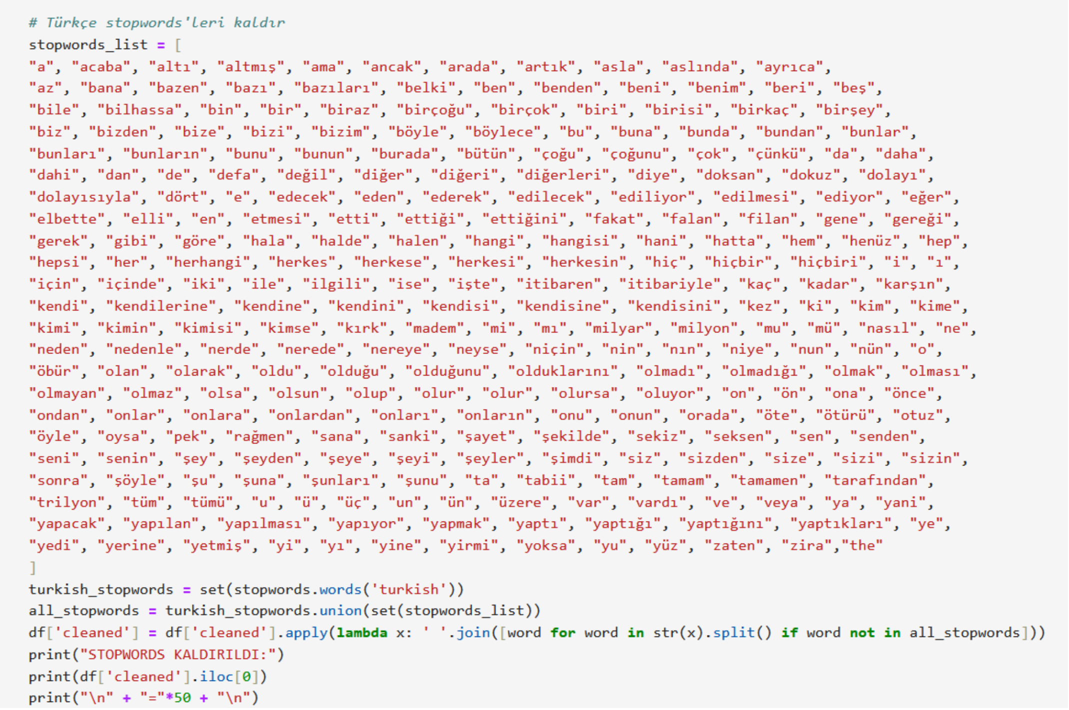
Removal of Numbers: Numbers in the text are removed to keep the focus of the analysis on the words. The relevant code block is detailed in Figure 7.



```
df['cleaned'] = df['cleaned'].apply(lambda x: re.sub(r'\d+', '', str(x)))
print("SAYILAR KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

Figure 7. Data Preprocessing – Removal of Numbers

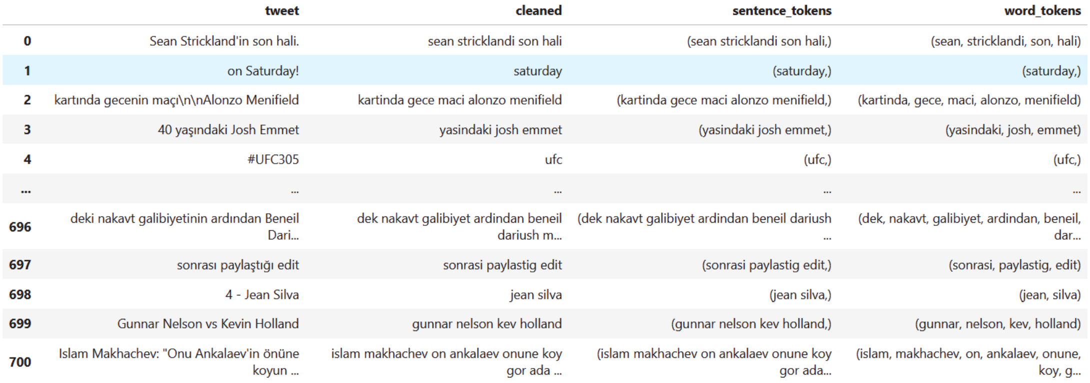
Removal of Stopwords: Frequently used stopwords such as "ve", "ile", "ama", "gibi", which generally do not have any meaning on their own, are removed from the text using a list specially prepared for Turkish. The relevant code block is given in detail in Figure 8.



```python
# Türkçe stopwords'leri kaldır
stopwords_list = [
"a", "acaba", "altı", "altmış", "ama", "ancak", "arada", "artık", "asla", "aslında", "ayrıca",
"az", "bana", "bazen", "bazı", "bazıları", "belki", "ben", "benden", "beni", "benim", "beri", "beş",
"bile", "bilhassa", "bin", "bir", "biraz", "birçoğu", "birçok", "biri", "birisi", "birkaç", "birşey",
"biz", "bizden", "bize", "bizi", "bizim", "böyle", "böylece", "bu", "buna", "bunda", "bundan", "bunlar",
"bunları", "bunların", "bunu", "bunun", "burada", "bütün", "çoğu", "çoğunu", "çok", "çünkü", "da", "daha",
"dahi", "dan", "de", "defa", "değil", "diğer", "diğeri", "diğerleri", "diye", "doksan", "dokuz", "dolayı",
"dolayısıyla", "dört", "e", "edecek", "eden", "ederek", "edilecek", "ediliyor", "edilmesi", "ediyor", "eğer",
"elbette", "elli", "en", "etmesi", "etti", "ettiği", "ettiğini", "fakat", "falan", "filan", "gene", "gereği",
"gerek", "gibi", "göre", "hala", "halde", "halen", "hangi", "hangisi", "hani", "hatta", "hem", "henüz", "hep",
"hepsi", "her", "herhangi", "herkes", "herkese", "herkesi", "herkesin", "hiç", "hiçbir", "hiçbiri", "i", "ı",
"için", "içinde", "iki", "ile", "ilgili", "ise", "işte", "itibaren", "itibariyle", "kaç", "kadar", "karşın",
"kendi", "kendilerine", "kendine", "kendini", "kendisi", "kendisine", "kendisini", "kez", "ki", "kim", "kime",
"kimi", "kimin", "kimisi", "kimse", "kırk", "madem", "mi", "mı", "milyar", "milyon", "mu", "mü", "nasıl", "ne",
"neden", "nedenle", "nerde", "nerede", "nereye", "neyse", "niçin", "nin", "nın", "niye", "nun", "nün", "o",
"öbür", "olan", "olarak", "oldu", "olduğu", "olduğunu", "olduklarını", "olmadı", "olmadığı", "olmak", "olması",
"olmayan", "olmaz", "olsa", "olsun", "olup", "olur", "olur", "olursa", "oluyor", "on", "ön", "ona", "önce",
"ondan", "onlar", "onlara", "onlardan", "onları", "onların", "onu", "onun", "orada", "öte", "ötürü", "otuz",
"öyle", "oysa", "pek", "rağmen", "sana", "sanki", "şayet", "şekilde", "sekiz", "seksen", "sen", "senden",
"seni", "senin", "şey", "şeyden", "şeye", "şeyi", "şeyler", "şimdi", "siz", "sizden", "size", "sizi", "sizin",
"sonra", "şöyle", "şu", "şuna", "şunları", "şunu", "ta", "tabii", "tam", "tamam", "tamamen", "tarafından",
"trilyon", "tüm", "tümü", "u", "ü", "üç", "un", "ün", "üzere", "var", "vardı", "ve", "veya", "ya", "yani",
"yapacak", "yapılan", "yapılması", "yapıyor", "yapmak", "yaptı", "yaptığı", "yaptığını", "yaptıkları", "ye",
"yedi", "yerine", "yetmiş", "yi", "yı", "yine", "yirmi", "yoksa", "yu", "yüz", "zaten", "zira","the"
]
turkish_stopwords = set(stopwords.words('turkish'))
all_stopwords = turkish_stopwords.union(set(stopwords_list))
df['cleaned'] = df['cleaned'].apply(lambda x: ' '.join([word for word in str(x).split() if word not in all_stopwords]))
print("STOPWORDS KALDIRILDI:")
print(df['cleaned'].iloc[0])
print("\n" + "="*50 + "\n")
```

Figure 8. Data Preprocessing – Removal of Stopwords

As a result of these steps, the clean text obtained is ready for analysis. The cleaned version of the texts after the code blocks used is given in Figure 9.



	tweet	cleaned	sentence_tokens	word_tokens
0	Sean Strickland'in son hali.	sean stricklandi son hali	(sean stricklandi son hali,)	(sean, stricklandi, son, hali)
1	on Saturday!	saturday	(saturday,)	(saturday,)
2	kartında gecenin maçı\n\nAlonzo Menifield	kartinda gece maci alonzo menifield	(kartinda gece maci alonzo menifield,)	(kartinda, gece, maci, alonzo, menifield)
3	40 yaşındaki Josh Emmet	yasindaki josh emmet	(yasindaki josh emmet,)	(yasindaki, josh, emmet)
4	#UFC305	ufc	(ufc,)	(ufc,)
...	...	...	...	...
696	deki nakavt galibiyetinin ardından Beneil Dari...	dek nakavt galibiyet ardindan beneil dariush m...	(dek nakavt galibiyet ardindan beneil dariush ...	(dek, nakavt, galibiyet, ardindan, beneil, dar...
697	sonrası paylaştığı edit	sonrasi paylastig edit	(sonrasi paylastig edit,)	(sonrasi, paylastig, edit)
698	4 - Jean Silva	jean silva	(jean silva,)	(jean, silva)
699	Gunnar Nelson vs Kevin Holland	gunnar nelson kev holland	(gunnar nelson kev holland,)	(gunnar, nelson, kev, holland)
700	Islam Makhachev: "Onu Ankalaev'in önüne koyun ...	islam makhachev on ankalaev onune koy gor ada ...	(islam makhachev on ankalaev onune koy gor ada...	(islam, makhachev, on, ankalaev, onune, koy, g...

Figure 9. Cleaned Text

3.3. *Analysis module*

The analysis module given in Figure 10 consists of preprocessing, sentiment analysis, word cloud and LDA topic modeling modules. Each analysis type has its own optimized processing steps. The results are presented to the user in visualization and statistical format.

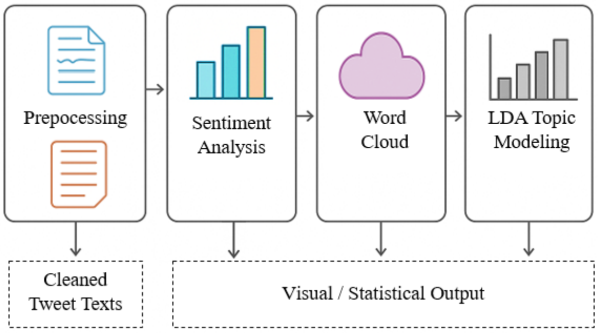


Figure 10. Analysis Module

Three different analysis techniques are applied on the preprocessed data:

Sentiment Analysis: A technique used to automatically detect and classify emotional content in texts. This method tries to determine whether a text carries a positive, negative or neutral sentiment. Sentiment analysis is usually used on unstructured data such as social media posts, product reviews, customer feedback[20]. In this study, sentiment analysis will analyze the general trend of the user's tweets by estimating their emotional state. The emotional load (positive, negative, neutral) of each tweet is classified. In the sentiment analysis phase, a pre-trained BERT-based deep learning model named "savasy/bert-base-turkish-sentiment-cased" from the HuggingFace platform was used. This model was specifically trained to classify sentiment in Turkish texts and fine-tuned on a large corpus of Turkish texts consisting of Turkish social media data, product reviews, and news articles. The model classifies the input texts into three categories: positive, negative, and neutral. Furthermore, when the model's confidence score falls below a certain threshold (60%), the relevant tweet is labeled as neutral to prevent ambiguous classifications. As a result of the classification for all tweets, a distribution is obtained that shows the general sentiment tendency of the user.

Word Cloud: It provides quick information about the general content of the text by visualizing the most frequently used words in the text. Word clouds provide quick understanding of the main themes and focal points of the text by visualizing the most frequently used words in a text. This is useful for summarizing long texts, analyzing content trends, and understanding patterns in social media conversations [21]. In this study, it is aimed to obtain information about tweets at a glance by performing word cloud analysis on tweets. It is used to visualize word frequencies in texts. The most frequently occurring words in pre-processed texts are determined and visualized as a "cloud" by enlarging them in proportion to their frequency. This provides a quick view of the user's language use and interests.

Topic Modeling: It is a technique that automatically detects hidden themes or topics in large text volumes. In this method, each document is viewed as a mixture of various topics and each topic is represented by specific words. Topic modeling facilitates thematic analysis, especially in large data sets, allowing texts to be organized and interpreted [22]. In this study, topic modeling was used to reveal hidden topics in tweets. Latent Dirichlet Allocation (LDA) model, a probabilistic based topic modeling method, is used to reveal thematic structures hidden in the text collection of user tweets. LDA assumes that each tweet is a mixture of one or more topics and each topic is represented by a distribution of certain words. As a result of the analysis, the topics that the user mentions the most and the keywords that characterize these topics are determined. The number of topics to be used in LDA topic distribution analysis is determined by the user via the webpage. A default value of 5 topics is assigned, but users can change this value as needed depending on the size and content diversity of their dataset. This approach provides flexibility for different data packages and research data and allows for the inclusion of personal domain information analyses. In studies, considering the short text structure and topic diversity of the X data, meaningful results have always been observed with topic numbers between 3 and 6.

3.4. Web module

The web interface developed with the Flask framework offers the following functions to the user:

- X data extraction
- Analysis type selection
- Visual and textual presentation of results

3.5. Technologies and libraries used

The technologies and libraries used in this study are listed below.

- *Programming Language:* Python (3.8+) was preferred, which provides rich library support for data science and web development.
- *Web Framework:* Flask was used due to its simple and extensible structure. Form management and authentication were provided with plugins such as Flask-Login and Flask-WTF.
- *Data Fetch:* Twikit library, which can fetch data without requiring X API, is used with cookie-based authentication support.
- *NLP Libraries:* Transformers were preferred for BERT-based sentiment analysis, NLTK for basic text processing, Wordcloud for visualization, Gensim for LDA, Regex for text cleaning.
- *Visualization Libraries:* Matplotlib and Seaborn were preferred for graphics, PyLDAvis was preferred for interactive LDA results.

4. Application

Within the scope of this study, a web interface has been developed that allows users to check tweets belonging to anyone they want and display that person's emotional state, topic modeling, and a word cloud consisting of the most frequently used words. The developed web-based application analyzes tweets belonging to any X account of users and performs various natural language processing-based analyses such as sentiment analysis, topic modeling, and word cloud. Thanks to the user-friendly interface, people without technical knowledge can easily perform these analyses.

The user starts the process by entering the username of the account they want to analyze and the number of tweets they want to pull on the main screen of the application. The system collects tweets from the relevant account in line with this information and saves them in JSON format. Then, the tweet data is passed through various pre-processing steps. The cleaned texts are sent to the analysis module. All analyzes are visualized through a user-friendly interface and the user can download the analysis results as a CSV file or as images if they wish. The main purpose of the application is to enable even users without technical knowledge to quickly analyze the content structure, emotional state and interests of an X account with a simple login.

The data flow in the system shown in Figure 11 generally follows the following steps:

- The user enters the target account and tweet count on the login screen
- Tweets are pulled, saved as JSON, and analysis is started
- The cleaned text is sent to the analysis modules (sentiment, word cloud, LDA).
- Analysis results are produced as graphics and statistics.
- CSV and visual output options are offered to the user.

The pseudocode of the algorithm containing the applied process steps is given below.

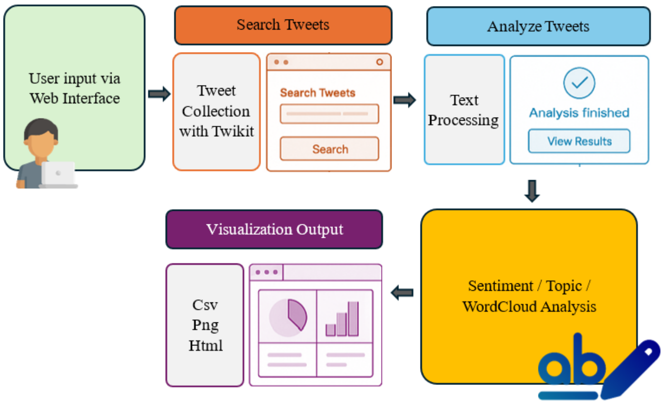


Figure 11. End-to-End Analysis System

[Algorithm 1]

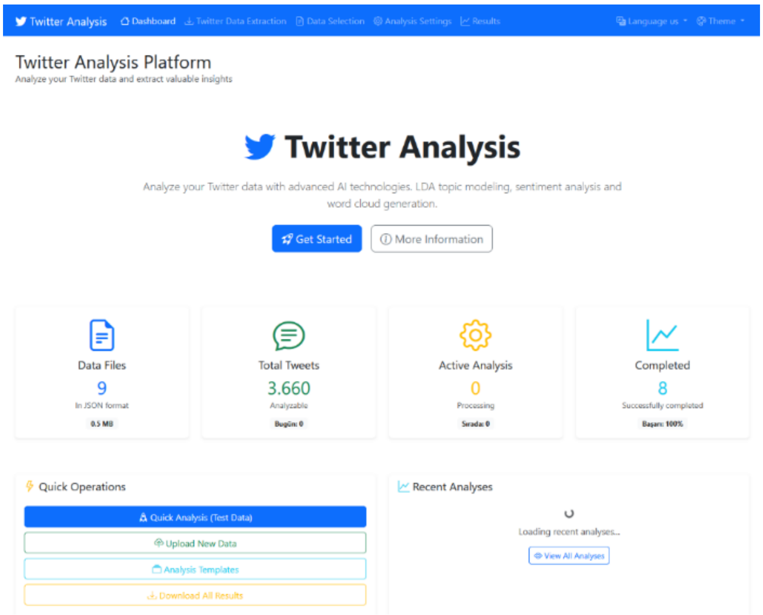
```
Function analyze_user_from_X(username):
    Step 1: Tweet Collection(Web Scraping)
            Tweets ← Take public tweets(username) as json
    Step 2: Data processing
            For each tweet in Tweets:
                Convert to lower case
                Clean URL, mention, hashtag, punctuation, special characters, html
tags
                Remove Turkish stopwords, punctuation and numbers
            Take all tweets as document
    Step 3: Analyze Tweets with LDA
            Sentiment Analysis
            Topic Analysis
            Word Cloud Analysis
    Step 4: Visualization Output
            Topic distributions
            Most frequent thematic words
            Visualization data (for graph and word cloud)
            Download results (csv, png, html)
end function
```

4.1. Web interface

The input image of the designed web interface is in Figure 12.

The data extraction system for the data collection phase is given in Figure 13. At this stage, the Data Extraction process step is started by receiving the username and the number of tweets that are wanted to be extracted from outside. The process steps applied in the background are listed below.

- Access to X via cookie,
- Big data extraction with pagination,
- Storing elements such as tweet text, date, interaction counts as JSON
- Error management mechanism



Analyze your Twitter data and extract valuable insights

Analyze your Twitter data with advanced AI technologies. LDA topic modeling, sentiment analysis and word cloud generation.

Data Files: 9, In JSON format, 0.5 MB

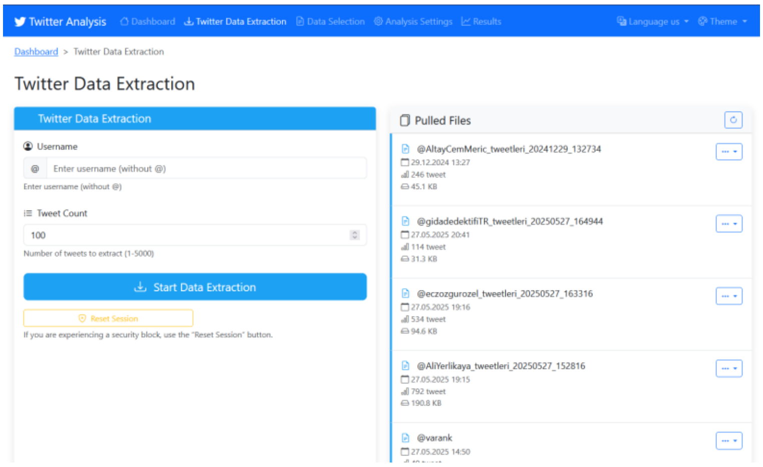
Total Tweets: 3.660, Analyzable, Bugün: 0

Active Analysis: 0, Processing, Sırada: 0

Completed: 8, Successfully completed, Başarı: 100%

Loading recent analyses...

Figure 12. Web Interface User Screen



Username: @ Enter username (without @)

Enter username (without @)

Tweet Count: 100

Number of tweets to extract (1-5000)

If you are experiencing a security block, use the "Reset Session" button.

@AltayCemMeric_tweetleri_20241229_132734
29.12.2024 13:27
246 tweet
45.1 KB

@gidadedektifiTR_tweetleri_20250527_164944
27.05.2025 20:41
114 tweet
31.3 KB

@eczozgurozel_tweetleri_20250527_163316
27.05.2025 19:16
534 tweet
94.6 KB

@AliYerlikaya_tweetleri_20250527_152816
27.05.2025 19:15
792 tweet
190.8 KB

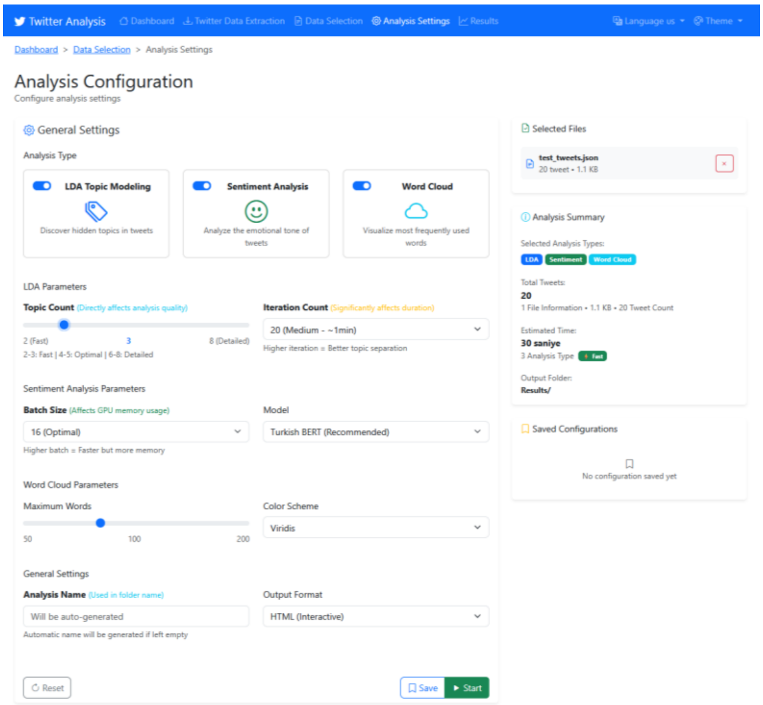
@varank
27.05.2025 14:50

Figure 13. Data Extraction System

After the data loading process steps are completed, data pre-processing steps are applied. The operations performed at this stage are given in detail below.

- HTML, URL, mention, hashtag, emoji and numbers are cleaned with Regex
- Turkish-specific normalization operations are performed
- Lower case conversion, space deletion, letter repetition reduction
- Stopword cleaning, word length and frequency filtering

Then, the user is directed to the Analysis Configuration page shown in Figure 14 to perform the sentiment analysis. The developed web interface is given in Figure 14. At this stage, the analysis process is started by receiving from the user the LDA parameters in the form of topic count, iteration count, batch size, BERT-based model; word cloud parameters in the form of maximum words, color scheme (seaborn, viridis) selection for sentiment distribution graphs, and general adjustment parameters in the form of analysis name, output format (csv, html), and confidence score threshold setting. The completion stage of the analysis process is prepared in a user-friendly way and what is done in each stage is presented to the user as in Figure 15.



Analysis Configuration
Configure analysis settings

General Settings

Analysis Type

LDA Topic Modeling
Discover hidden topics in tweets

Sentiment Analysis
Analyze the emotional tone of tweets

Word Cloud
Visualize most frequently used words

LDA Parameters

Topic Count (Directly affects analysis quality)
2 (Fast) 3 8 (Detailed)
2-3: Fast | 4-5: Optimal | 6-8: Detailed

Iteration Count (Significantly affects duration)
20 (Medium - ~1min)
Higher iteration = Better topic separation

Sentiment Analysis Parameters

Batch Size (Affects GPU memory usage)
16 (Optimal)
Higher batch = Faster but more memory

Model
Turkish BERT (Recommended)

Word Cloud Parameters

Maximum Words
50 100 200

Color Scheme
Viridis

General Settings

Analysis Name (Used in folder name)
Will be auto-generated
Automatic name will be generated if left empty

Output Format
HTML (Interactive)

Selected Files
test_tweets.json
20 tweet • 1.1 KB

Analysis Summary

Selected Analysis Types:
LDA Sentiment Word Cloud

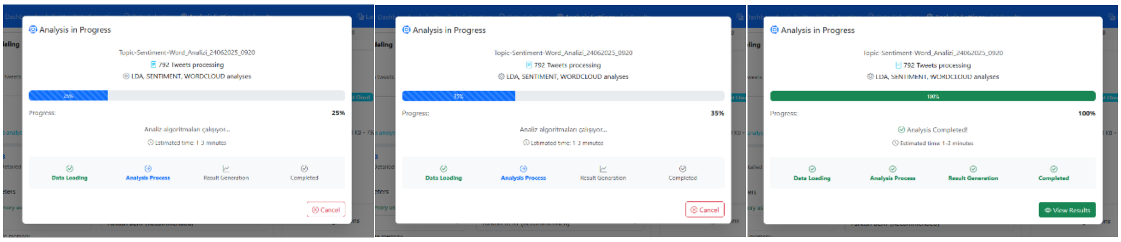
Total Tweets:
20
1 File Information • 1.1 KB • 20 Tweet Count

Estimated Time:
30 saniye
3 Analysis Type Fast

Output Folder:
Results/

Saved Configurations
No configuration saved yet

Figure 14. Analysis Configuration


Analysis in Progress
Topic-Sentiment-Word_Analizi_24062025_0920
792 Tweets processing
LDA, SENTIMENT, WORDCLOUD analyses
25%
Progress: 25%
Analiz algoritmaları çalışıyor...
Estimated time: 1-3 minutes

Analysis in Progress
Topic-Sentiment-Word_Analizi_24062025_0920
792 Tweets processing
LDA, SENTIMENT, WORDCLOUD analyses
35%
Progress: 35%
Analiz algoritmaları çalışıyor...
Estimated time: 1-3 minutes

Analysis in Progress
Topic-Sentiment-Word_Analizi_24062025_0920
792 Tweets processing
LDA, SENTIMENT, WORDCLOUD analyses
100%
Progress: 100%
Analysis Completed!
Estimated time: 1-3 minutes

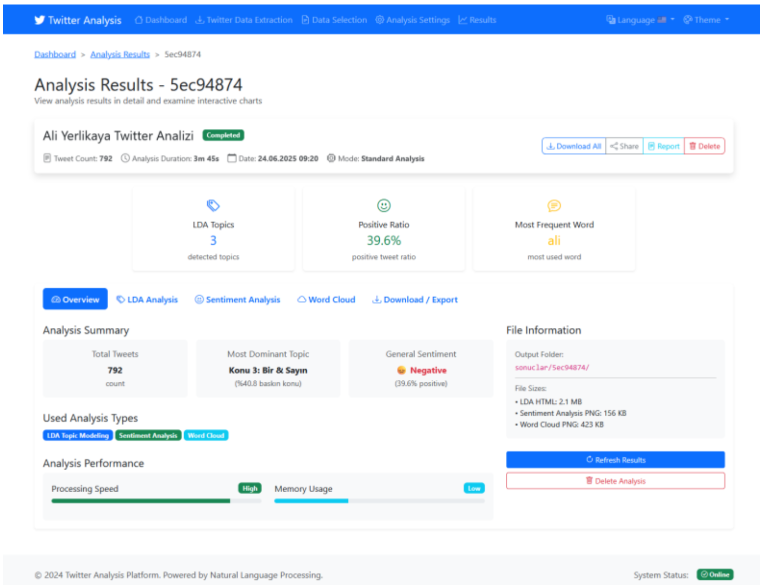
Figure 15. Analysis completion stage

In order to see the results of the analysis, they are transferred to an interface as shown in Figure 16. Here, there are LDA analysis results, Sentiment Analysis, Word Cloud and Download/Export sections. In the general view, the analysis performances are presented to the user along with information including the total number of tweets, how many topics there are and the general sentiment.

The interface prepared for LDA analysis is given in Figure 17. The process steps applied in this section are;

- LDA model training with Gensim,

- Summary extraction with the most meaningful words per topic,
- Interactive visual with pyLDAvis,
- Word cloud, scatter plot and heat map for each topic.



Dashboard > Analysis Results > 5ec94874

Analysis Results - 5ec94874

View analysis results in detail and examine interactive charts

Ali Yerlikaya Twitter Analizi

Tweet Count: 792 Analysis Duration: 3m 45s Date: 24.06.2025 09:20 Mode: Standard Analysis

LDA Topics 3 detected topics

Positive Ratio 39.6% positive tweet ratio

Most Frequent Word ali most used word

Analysis Summary

Total Tweets	Most Dominant Topic	General Sentiment
792 count	Konu 3: Bir & Sayın (%40.8 baskın konu)	Negative (39.6% positive)

Used Analysis Types

Analysis Performance

Processing Speed

Memory Usage

File Information

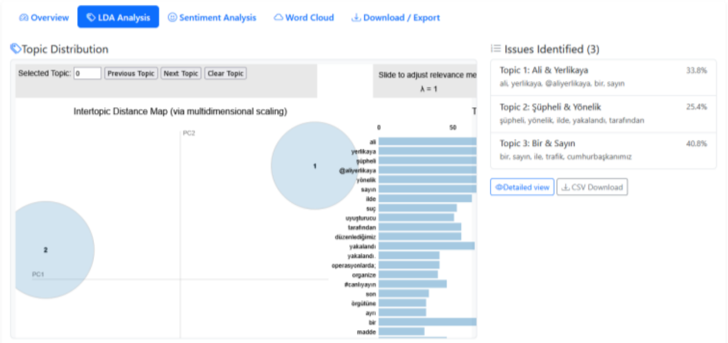
Output Folder:
`sonuclar/5ec94874/`

File Sizes:
- LDA HTML: 2.1 MB
- Sentiment Analysis PNG: 156 KB
- Word Cloud PNG: 423 KB

© 2024 Twitter Analysis Platform. Powered by Natural Language Processing.

System Status:

Figure 16. Analysis result screen



Topic Distribution

Issues Identified (3)

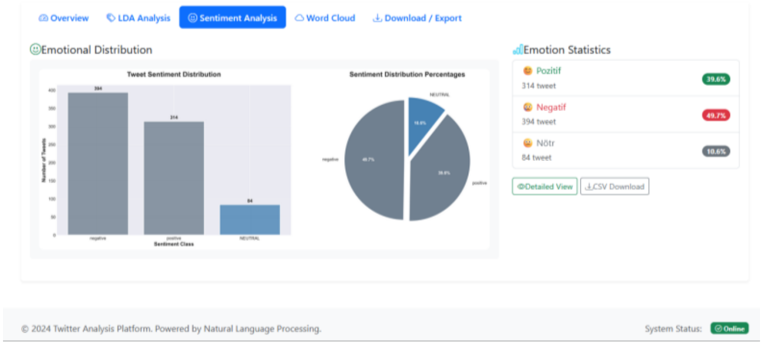
Topic 1: Ali & Yerlikaya ali, yerlikaya, @aliyerlikaya, bir, sayın	33.8%
Topic 2: Şüpheli & Yönelik şüpheli, yönelik, ilde, yakalandı, tarafından	25.4%
Topic 3: Bir & Sayın bir, sayın, ile, trafik, cumhurbaşkanımız	40.8%

Figure 17. LDA Analysis interface

The interface prepared for sentiment analysis is given in Figure 18. In this section, sentiment distributions are grouped under 3 different headings: positive, negative and neutral. Visuals of how many tweets there are from which sentiment and their percentages are given with bar and pie charts.

The interface prepared for the word cloud is given in Figure 19. The process steps applied in this section are;

- Extracting word frequencies from cleaned texts,
- Determining the most frequently occurring words with collections.counter,
- PNG visual output with WordCloud,
- CSV file containing word frequencies.



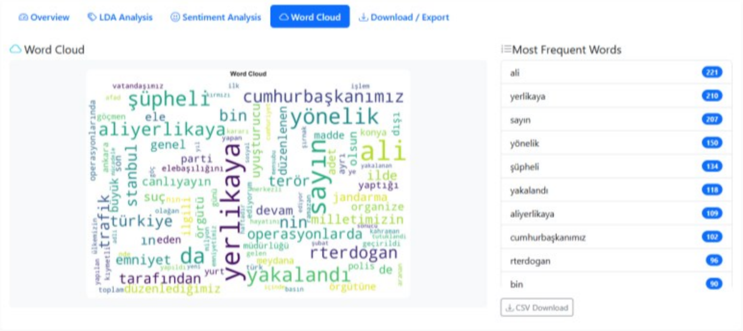
Pozitif 314 tweet 39.6%

Negatif 394 tweet 49.7%

Nötr 84 tweet 10.6%

© 2024 Twitter Analysis Platform. Powered by Natural Language Processing.

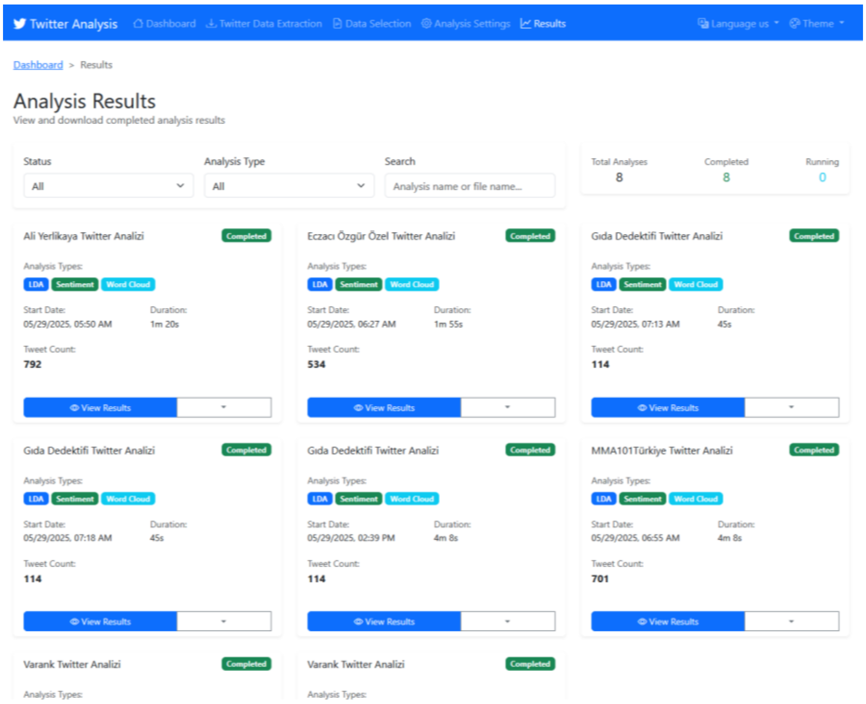
Figure 18. Sentiment Analysis interface



ali	221
yerlikaya	210
sayın	207
yönelik	150
şüpheli	134
yakalandı	118
aliyerlikaya	109
cumhurbaşkanımız	102
rterdogan	96
bin	90

Figure 19. Word cloud interface

The general interface of the X accounts that have been previously analyzed is given in Figure 20. Information such as how long the operations took, the date they were performed, and the total number of tweets are provided, as well as details of the results obtained are also provided to the user.



Analysis Results

View and download completed analysis results

Status: All

Analysis Type: All

Search: Analysis name or file name...

Total Analyses 8 | Completed 8 | Running 0

Ali Yerlikaya Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 05:50 AM
Duration: 1m 20s
Tweet Count: 792

Eczacı Özgür Özel Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 06:27 AM
Duration: 1m 55s
Tweet Count: 534

Gıda Dedektifi Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 07:13 AM
Duration: 45s
Tweet Count: 114

Gıda Dedektifi Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 07:18 AM
Duration: 45s
Tweet Count: 114

Gıda Dedektifi Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 02:39 PM
Duration: 4m 8s
Tweet Count: 114

MMA101Türkiye Twitter Analizi — Completed
Analysis Types: LDA, Sentiment, Word Cloud
Start Date: 05/29/2025, 06:55 AM
Duration: 4m 8s
Tweet Count: 701

Varank Twitter Analizi — Completed
Analysis Types:

Varank Twitter Analizi — Completed
Analysis Types:

Figure 20. Analysis Result Interface

4.2. Sentiment analysis

In this section, the results of the analysis conducted on a sample user X are presented and interpreted in order to demonstrate the functionality of the developed system. As a result of the sentiment analysis conducted on the tweets of the user named @gidadedektifiTR, it was determined that 8.8% of the tweets were positive, 88.6% were negative and 2.6% were neutral. The sentiment analysis results obtained are given in Figure 21.

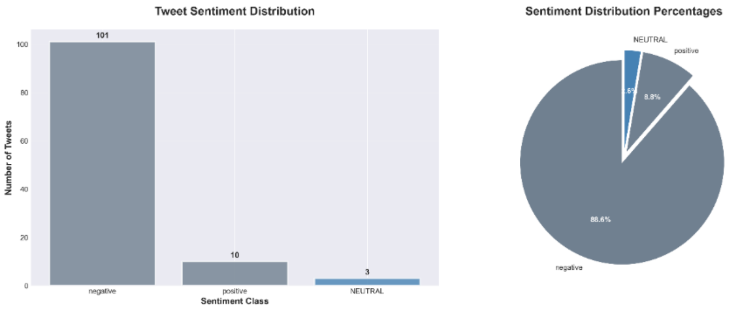


Figure 21. Sentiment analysis results

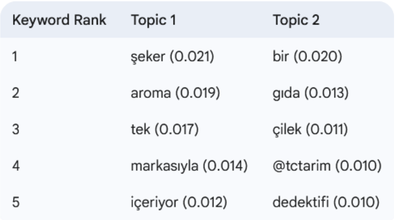
When the tweets classified as negative were examined, it was seen that they generally focused on topics such as the content of food products, deceptive packaging, additives, palm oil use, lack of inspections and practices that mislead the public. The tweets critically examine the content details of the products and aim to raise consumer awareness.

Positive tweets, on the other hand, are more related to celebrations of success, promotion of informative social media accounts and some social reactions. For example, posts about sports achievements such as Fenerbahçe Beko's championship and recommended accounts to follow contain positive emotions.

These findings suggest that the user is sensitive to food safety, consumer rights and conscious consumption issues. It is also observed that he values positive content such as social achievements and information sharing.

Topic Modeling Results

As a result of LDA analysis, 2 main topics that were prominent in the user's tweets were determined. These topics and the words that best represent these topics are shown in Figure 22.



Keyword Rank	Topic 1	Topic 2
1	şeker (0.021)	bir (0.020)
2	aroma (0.019)	gıda (0.013)
3	tek (0.017)	çilek (0.011)
4	markasıyla (0.014)	@tctarim (0.010)
5	içeriyor (0.012)	dedektifi (0.010)

Figure 22. LDA topic table

The theme determined as "Topic 1" can be called "Content and Features of Food Products". Words such as sugar, aroma, milk, oil, contains, product and brand collected under this theme show that the posts are mostly about the label information, nutritional values and ingredients of the products. This reflects the tendency of consumers to understand and analyze what is in the products.

"Topic 2" can be called "Food Safety and Agricultural Inspection". The fact that this theme is represented by words such as @tctarim (Ministry of Agriculture and Forestry), food detective, detection, agriculture, cattle and pig reveals that the conversations are about the reliability, inspection and source of food products. The fact that official institutions and popular market chains are also involved in these discussions shows that the topic has a public dimension and is closely related to consumer trust.

The distribution graph of these topics is given in detail in Figure 23.

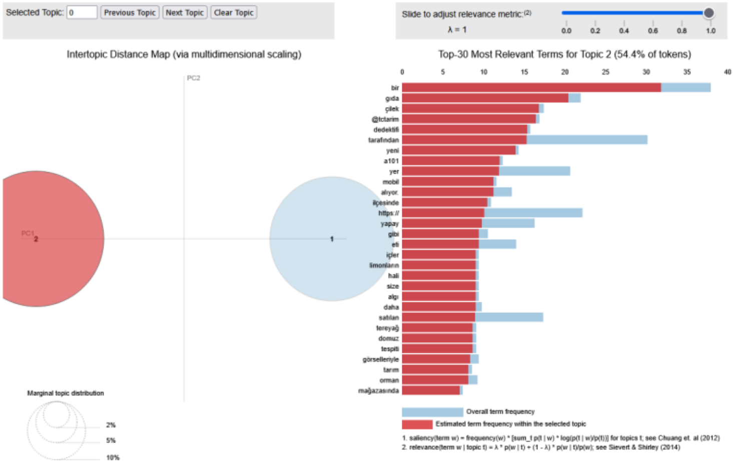

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (54.4% of tokens)
0
5
10
15
20
25
30
35
40
bir
gıda
çilek
@tctarim
dedektifi
tarafından
yeni
a101
yer
mobil
alıyor.
ilçesinde
https://
yapay
gibi
eti
içler
limonların
hali
size
algı
daha
satılan
tereyağ
domuz
tespiti
görselleriyle
tarım
orman
mağazasında
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 23. Topic distribution graph

Word Cloud Analysis

The word cloud created from all tweets of the user visually summarizes the most frequently used words. The analysis result obtained is given in Figure 24. The size of words such as food, detective, product, in its content and brand in the word cloud shows that the main focus of the posts in this dataset are food products, their content analysis and reliability checks. This visual directly supports the previous topic analysis results and allows understanding of the general characteristics of the texts at a glance.

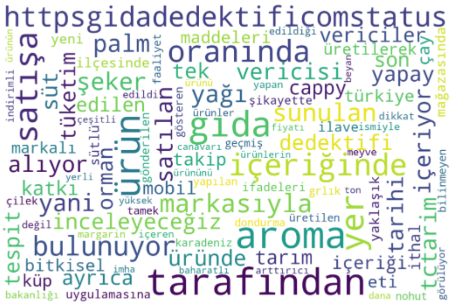


Figure 24. Word Cloud Analysis Results

The word analysis results are given in Figure 25. When the results are examined, graphs containing the most frequently used words, word frequencies and word lengths are given.

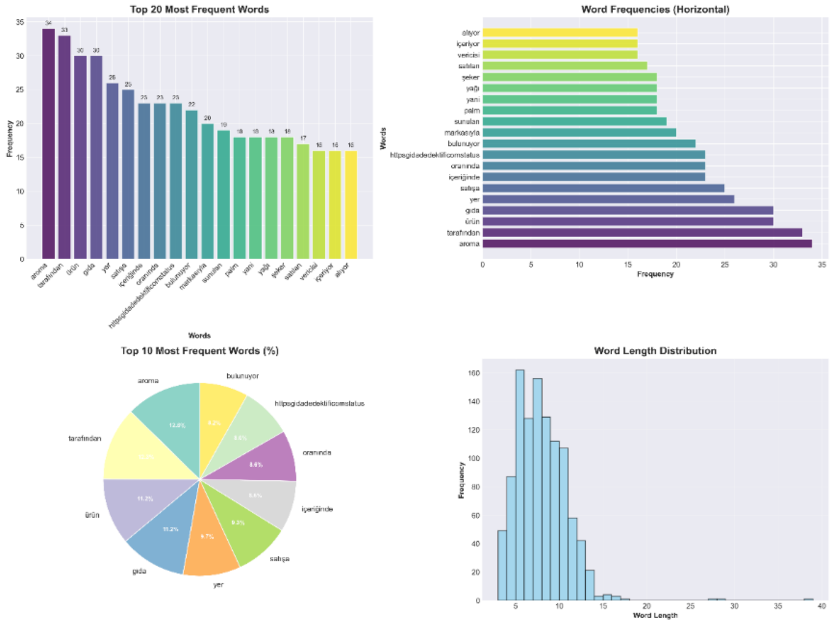


Figure 25. Word Analysis Results

Table 1 shows sentiment analysis using tweets from 10 different individuals. Tweets published by each individual were examined and classified as positive, negative, and neutral, and the total number of tweets and the percentages of these classes are detailed.

When the tweets of the same users are examined in Table 2, the total number of tweets examined, the number of topics these tweets are related to, the average tweet length, the number of characters in their longest tweets, the number of words, and the number of hashtags they used are given.

Table 3 shows the topics related to the tweets published and the rate of each topic by these 10 different users.

Table 1. Sentiment analysis for 10 different users

	Tweet Count			Ratio(%)		
	Positive	Negative	Neutral	Positive	Negative	Neutral
@user1	679	350	135	58	30	12
@user2	286	429	77	36	54	10
@user3	263	460	75	27	65	8
@user4	181	20	12	85	9	6
@user5	281	149	18	63	33	4
@user6	323	173	38	60	32	7
@user7	455	167	68	66	24	10
@user8	277	530	62	32	61	7
@user9	349	346	108	43.5	43.1	13.4
@user10	310	514	116	33	55	12

Table 2. Detailed tweet analysis for 10 different users

User	Processed Tweets	Number of Topics	Average Tweet Length	Longest Tweet	Unique Word Count	Total Hashtags
@user1	1164	3	2018	450	5560	461
@user2	792	2	2150	293	4419	145
@user3	978	2	1889	556	8751	11
@user4	213	2	1992	371	950	0
@user5	448	3	1581	595	3579	5
@user6	534	3	1577	286	3753	58
@user7	690	2	601	277	2395	27
@user8	869	3	1323	299	5647	4
@user9	803	3	1062	280	4086	235
@user10	940	2	1298	295	4658	5

Table 3. Topic distribution and proportions for 10 different users

	Topic Label	Topic Proportion
@user1	Deprem Yardımı ve Destek Faaliyetleri	0.335
	Teşekkür ve Minnettarlık	0.343
	Genel İletişim ve Bağış Çağrıları	0.322
@user2	Siyaset ve Resmi Açıklamalar	0.543
	Suç Operasyonları ve Güvenlik	0.457
@user3	Genel Konuşma ve Sorular	0.724
	Youtube içeriği ve Yaratıcı	0.276
@user4	Spor Ligleri, Turnuvalar ve Tahminler	0.417
	Oyuncu Gündemi	0.583
@user5	Teknik Haberler ve Windows Sorunları	0.268
	Platform İletişimi ve Paylaşımlar	0.346
	Genel Sohbet ve İçerik Platformları	0.385
@user6	Kamuoyu ile İletişim ve Sorulara Yanıtlar	0.360
	Parti Faaliyetleri ve Anma Törenleri	0.379
	Genel Açıklamalar ve Ziyaretler	0.261
@user7	Dövüş Sonuçları ve Genel Tartışmalar	0.32
	UFC Dövüşleri ve Maç Gündemi	0.68
@user8	Yangın Müdahalesi ve Devam Eden Çabalar	0.313
	Yangınlara Siyasi ve İdari Müdahale	0.408
	Sosyal Medya, Siyaset ve Ekonomi	0.279
@user9	Teknoloji, Oyun ve Ürün Beklentileri	0.283
	Yapay Zeka ve Ugyulamaları	0.494
	Donanım ve Yazılım Güncellemeleri Duyuruları	0.223
@user10	Teknik Direktör ve Maç Yönetimi	0.56
	Futbol Transferleri ve Anlaşmalar	0.44

5. Conclusion and future works

In this study, a web system that can perform user-oriented text analysis on X data has been successfully developed and presented. The system performs data collection with web scraping, data preprocessing specific to the Turkish language, sentiment analysis, topic modeling and word cloud analysis steps in an integrated manner. The developed application, unlike general trend analyses in the literature, has the potential to provide deeper and more personalized insights by focusing on the individual user. It is seen that most of the literature studies perform sentiment analysis using a specific topic or specific keywords.

The obtained results show that the system achieves its goal by successfully revealing a user's sentiment tendencies, topics of interest and frequently used expressions. This tool can be a valuable analysis source for social scientists, marketers or individual enthusiasts.

The proposed system has some limitations. First, the system can only analyze publicly shared tweets and cannot access content from private or protected accounts. This limits its ability to represent the entire X user base. Furthermore, since the sentiment analysis model used is trained for Turkish texts, the system exhibits language dependency and requires the integration of models trained in the relevant language to analyze tweets in other languages.

In future studies, the system can be developed in the following directions:

- Time Series Analysis: A module that analyzes the changes in user sentiment and subject tendencies over time can be added.
- Comparative Analysis: A feature that compares multiple user profiles to reveal similarities and differences can be developed.
- Platform Support: The system can also be enabled to extract and analyze data from other social media platforms (e.g. Reddit, Instagram).

This study demonstrates the potential of user-centered social media analysis and provides a basis for future research in this area.

Acknowledgements

We would like to thank Karamanoğlu Mehmetbey University for providing support during this study. This work was derived from M.Y.B.'s undergraduate senior project.

References

[1] Bilgin, M. and İ.F. Şentürk, Sentiment analysis of tweets based on document vectors using supervised learning and semi-supervised learning **Balıkesir Üniversitesi Fen Bilimleri Enstitüsü Dergisi**, 21, 2, 822-839, (2019).

[2] Vatambeti, R., et al., Twitter sentiment analysis on online food services based on elephant herd optimization with hybrid deep learning technique, **Cluster Computing**, 1-5, (2023).

[3] Rabadán-Martín, I., et al., Topic-based engagement analysis: Focusing on hotel industry Twitter accounts, **Tourism Management**, 106, 104981, (2025).

[4] Martín, M.S., F.-W. Chen, and P.A. Urbistondo, Application of the LDA model to identify topics in telemedicine conversations on the X social network, **BMC Health Services Research**, 25, 1, 369, (2025).

[5] Dewi, D.A. and T.B. Kurniawan, Exploring Financial Trends through Topic Modeling and Time-Series Analysis: A Clustering Approach Using Latent

Dirichlet Allocation (LDA) on Twitter Data, **Journal of Digital Society**, 1, 1, 91-108, (2025).

[6] Atılgan, K.Ö. and H. Yoğurtcu, Sentiment Analysis of Twitter Posts of Cargo Company Customers **Çağ Üniversitesi Sosyal Bilimler Dergisi**, 18, 1, 31-39, (2021).

[7] Güneş, Y. and M. Arıkan, X (Twitter) Sentiment Analysis Based on Hybrid Approach: An Application for Online Food Ordering, **Bilişim Teknolojileri Dergisi**, 18, 2, 143-167, (2025).

[8] Kim, J., D. Kim, and E. Park, I know your stance! Analyzing Twitter users' political stance on diverse perspectives, **Journal of Big Data**, 12, 1, 14, (2025).

[9] Kwak, H., et al. What is Twitter, a social network or a news media?, ***Proceedings of the 19th international conference on World wide web***, 591-600, (2010).

[10] Uzun, C., Twitter Mentions of February 6th Earthquake: An Emotional Language Analysis, **Doğal Afetler ve Çevre Dergisi**, 9, 2, 503–517, (2023).

[11] Kartal, Y., TOJDE Dergisi üzerinde LDA ile konu modelleme. 2017, Anadolu University (Turkey).

[12] Taşbaş Ustaoğlu, E., Conceptualization of Human Robot Interaction Through Word Cloud Analysis **Econder Uluslararası Akademik Dergi**, 3, 2, 221-239, (2019).

[13] Günyaktı, R.İ. and N. Bursa, Investigation of the Perception toward Health Workers and Teachers on Twitter Data via Sentiment Analysis in the Covid19 Pandemic, **Selçuk İletişim**, 15, 1, 264-285, (2022).

[14] Seren, N. and M.H. Altıntaş, Turizm İşletmelerinde Marka Kişiliğinin Duygu Analizi Yöntemiyle Belirlenmesi, **International Journal of Social Inquiry**, 16, 1, 229-254, (2023).

[15] İlhan, N. and D. Sağaltıcı, Sentiment Analysis in Twitter **Harran Üniversitesi Mühendislik Dergisi**, 5, 2, 146-156, (2020).

[16] Akgül, E.S., C. Ertano, and D. Banu, Sentiment analysis with Twitter, **Pamukkale University Journal of Engineering Sciences** 22, 2, 106-110, (2016).

[17] Aydın, G. and İ. Hallaç, Automatic Topic Detection on Turkish Text **Fırat Üniversitesi Mühendislik Bilimleri Dergisi**, 33, 2, 599-606, (2021).

[18] Carle, V., Web Scraping Using Machine Learning. 2020, KTH Royal Institute of Technology, Stockholm, Sweden. p. 1-4.

[19] Çetin, V. and O. Yıldız, A comprehensive review on data preprocessing techniques in data analysis, **Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi**, 28, 2, 299-312, (2022).

[20] Can, U. and B. Alatas, Review of Sentiment Analysis and Opinion Mining Algorithms **International Journal of Pure and Applied Sciences**, 3, 1, 75-111, (2017).

[21] Kabir, A.I., et al., The Power of Social Media Analytics: Text Analytics Based on Sentiment Analysis and Word Clouds on R, **Informatica Economica**, 22, 1, (2018).

[22] Jelodar, H., et al., Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey, **Multimedia tools and applications**, 78, 15169-15211, (2019).